基于大语言模型和多智能体的自动化政务回复参考生成系统*

方岢愿[1]，许珂维[1†]

(1. 香港科技大学（广州） 创新创业与公共政策学域, 广东 广州 510000)

摘 要：在数字治理时代，政务平台需要及时有效地回应公民的政务咨询。然而，现有的政务问答系统主要由人工回复为主，自动化处理算法的辅助有限，这难以高效处理大数据时代大量的市民政务咨询需求。因此在数字治理时代政务平台需要建立更有效、更智能的问答系统来回应市民的政务咨询。如今大语言模型 (LLMs) 有望助力政务平台以自动化有效的方式处理市民的政务咨询。LLMs 可以提高政务平台与市民互动的效率，为各种类型的市民咨询提供自然语言回复。然而，现有通用 LLMs 对政务领域特定的表达理解有限，暂时无法像平台工作人员做出有效的回复。本研究基于 LLMs 和政务咨询问答历史向量数据库，利用多智能体技术构建专门针对政务咨询的回复参考生成系统（GovLLM）。在输入新的市民咨询后，系统能够生成切实有效的示例答案，供平台工作人员处理市民咨询时进行参考。该系统表现出比基线模型更好的文本生成性能，有利于提高政务平台回复市民咨询时的效率和有效性。

关键词：电子政务；大语言模型；问答系统；信息系统

中图分类号：TP391

Automatic Government Response Reference Generation System Based on Large Language Models and Multi-Agent

Fang Keyuan[1], Xu Kewei[1†]

(1. The Hong Kong University of Science and Technology (Guangzhou), Innovation, Policy and Entrepreneurship Thrust, Guangdong Guangzhou 510000, China)

Abstract: In digital governance era, governmental platforms need to respond to citizens' governmental inquires in a timely and effective manner. However, the existing government Q&A system mainly relies on manual responses, with limited assistance from automatic algorithms, which makes it difficult to efficiently handle the large volume of citizens' government consultation needs in big data era. Therefore, in the era of digital governance, government platforms need to establish more effective and intelligent Q&A systems to respond to citizens' government consultations. Nowadays, large language models (LLMs) are expected to help government platforms handle citizens' government consultations in an automated and effective manner. LLMs can improve the efficiency of government platforms' interactions with citizens and provide natural language responses to various types of citizen consultations. However, existing general LLMs have limited understanding of specific expressions in the government field and are temporarily unable to make effective responses like platform staff. This study builds a response reference system specifically for citizens' governmental inquires based on LLMs and a historical vector database of government consultation questions and answers (GovLLM) by Muti-Agent systems. After inputting new citizens' inquires, the system is able to generate practical and effective example answers for platform staff to refer to when handling citizens' inquires. The system shows better text generation performance than the baseline model, which is conducive to improving the efficiency and effectiveness of government platforms in responding to citizens' inquires.

Key words: digital government; large language models; Q&A system; information system

0 引言

政务服务的数字化转型是提升国家治理能力的关键环节，需要依赖政务部门之间的服务和信息集成与共享[1]，将数据价值置于政务服务改革的核心[2]。随着数字化政务平台的广泛应用，市民进行在线政务咨询的需求量迅速增长，需求类型也更

基金项目：广东省哲学社会科学规划 2024 年度青年项目（GD24YGL37）；广州市 2024 年度基础与应用基础研究专题青年博士“启航”项目（No.SL2023A04J01836）；广州市校（院）企联合资助项目基础研究计划（2023A03J0165）；广州市教育局高校科研项目（2024312153）

作者简介：方岢愿（1998-），男，广东揭阳人，博士研究生，主要研究方向为自然语言处理，电子政务，数据挖掘；许珂维（1989-），男（通信作者），河南南阳人，助理教授，博导，博士，主要研究方向为公民参与，智慧城市，大数据分析，公共政策（coreyxu@hkust-gz.edu.cn）.

加多样化，给政务平台的响应能力带来了更大的挑战[3]。在此背景下，政务问答系统应运而生，成为数字化政务时代下回应市民查询的有效工具，促进政务平台与市民之间的互动和沟通[4]。智能化和高效能的政务问答系统能够满足市民多样化政务咨询需求的主要途径，也是市政部门了解市民需求的重要渠道。因此，构建智能和高效的政务问答系统对市民和市政部门都具有重要价值[5]。我国高度重视电子政务的发展，并提出建设“回应型政府”的总目标[6]，要求各级政务部门积极回应市民诉求并提供实质性解决方案以解决问题。因此，推进“回应型政府”建设是政策要求，也是提高治理效能的重要举措。

传统的问答系统主要通过三种方式进行构建[7]。第一种是基于信息检索的系统，它通过在文档集合中搜索短文本片段来回答问题[8]。第二种为知识匹配系统，它通过将自然语言问题映射到结构化数据库上的查询来响应用户查询[8]。第三种方法是应用机器学习和深度学习方法来训练模型学习问题和答案之间的映射关系[9]。虽然这三种方法可以在通用问答系统中取得相对较好的结果，但当它们应用于数字化政务平台时，均存在一定的局限性。首先是这些方法对政务服务领域的一些专业和固定表达的理解水平相对较低。其次是它们的输出回答经常偏离用户的原始意图。这些缺点在数字化政务平台尤为明显。市民的咨询数量庞大且类型丰富，特别是我国不同省市之间的政策差异显著[10]。这需要政务平台具备强大的历史数据检索能力和考虑不同地区的多样性而生成贴合实际的回复。当前的政务问答系统难以掌握满足以上需求，导致目前大部分的政务问答系统政仍是以人工回复为主[11]。这难以满足市民日益增长的政务咨询需求，因此本研究尝试应用最新的大语言模型（LLMs）来弥补这些研究空白。

近年来，LLMs 取得了重大进展，特别是在问答任务中应用广泛[12]。LLMs 具备在不同领域生成类似人类和专业语言的文本的能力，从根本上重塑了人工智能的格局[13]。现存研究已经成功应用 LLMs 来解决金融[14]和医学[15]等垂直领域的特定问题。然而，它们在政务领域的应用和研究相对较少，表明该领域具有巨大的研究潜力。虽然 LLMs 在政务领域具有较大的研究前景，但是其仍存在一定的问题需要解决。首先，LLMs 偶尔会生成事实不相符的内容[16]。其次，它们生成的文本会存在内容空洞和模板化，缺乏实用的有用信息的问题[17]。与其他领域问答系统不同的是，政务平台的回复需要严格符合客观事实和并且能切实有效地解决市民的问题，以此增强市民对市政部门的信任和满意度[18]。因此，本研究试图建立一个基于 LLMs 的政务回复参考生成系统来弥补该领域的研究空白。本研究的贡献可以总结如下：

a) 本研究以建设回应型政务平台为总目标，构建了一个基于 LLMs 和历史政务问答向量数据库的政务参考回复生成系统（GovLLM），相比于基线方法，模型表现提升了 5-11%，填补了 LLMs 在该领域的研究空白，显示了 LLMs 理解政务文本的强大能力。

b) 该系统生成的回复参考能够提供实际有效的解决方案，而不是空洞的模板回答。此外，该系统还引入历史向量数据库、智能体设计、位置过滤、问题类别分类、基于问题和回复属性的排序等模块，能够提高回复生成的效率，确保任务回复内容与咨询来源地区的客观事实保持一致，从而减少错误信息。这增强了系统的可用性和可靠性，有利于促进市民对政务平台的信任和满意度。

c) 该系统不仅在提高效率和文本质量上取得较大提升，而且也有利于降低人工劳动成本。传统的政务咨询平台依赖于人工回答，其相关的人工成本较高。该系统生成的回复参考能减轻了政务工作人员在处理公政务咨询上投入的时间和精力，有利于政务平台高效的运作。

1 相关研究

在大数据时代，建设回应型政府已成为各国进行政务服务优化的核心话题。现有研究表明，回应型政府需要政务部门及时且有效地响应市民需求，以及通过数字化手段实现政务平台与市民之间的良性互动[19]。也就是说，以市民需求为中心的服务是构建数字时代回应型政府的目标。与传统以政务部门自身为中心的政务服务相比，数字时代的回应型政府应专注于提供以公民为中心的政务服务，鼓励市民积极参与其中[19]。这要求政务部门在服务过程中需要更加注重市民体验和反馈，并应用创新数字技术从传统公共服务向智能服务转型[20]。然而，现有研究更多为实证和理论研究，以搭建回应型政府为主要研究目标的可落地的应用研究暂且缺乏。因此，本研究尝试结合最新的 LLMs 技术，搭建一个自动化政务服务回复参考的生成系统，以此作为搭建回应型政府的应用实例，为后续的相关研究提供参考。

在数字政务的应用领域，现有研究主要致力于搭建特定应用场景的基础政务问答系统。Georgios 等人搭建了一个用于护照政务场景的聊天机器人，解决了护照服务相关技术和公共管理方面的问题[21]。Schwarzer 等人设计了一个服务于德国柏林地区的数字政务问答系统，解决柏林市民在访问政府服务时面临的挑战[22]。Beltran 等人提出了一个问答系统与政府门户网站的集成系统架构，方便市民通过问答的方式触达他们所需要的政务服务[23]。Rodsawang 等人设计了一个应用于新冠疫情求助场景的政务聊天机器人，增强了政务部门对于新冠疫情的应急响应能力[24]。虽然这些研究取得了积极的结果，但大多数系统采用传统的问答系统搭建方法，存在以下问题：第一，现有应用研究由于数据集和系统搭建方法的限制，多数为区域级别的政务问答系统，其覆盖的地域范围有限。对于政策差异较大的大范围地区，相同的政务服务需求会因政策差异而产生不同的回复，而传统的系统搭建方法难以实现大范围地区的差异化政务问答。在人口高速流动，区域一体化，以及全国一张网的趋势下，之前的各个地区单独提供解决方案的方式亟需转变。 第二，传统的问答系统搭建系统因为模型复杂度和算法性能等原

因，其检索和生成速度较慢，在市民政务服务需求日益增长的大数据时代难以做到快速回复。最后，目前研究提出的政务问答系统多数停留于完成基础的问答功能和搭建基本的系统框架，而对于系统回复的有效性并未过多涉及和优化，这导致一些模板回复和无效回复的产生，这难以满足当下回应型政府的建设需求。因此，本研究将采用最新的 LLMs 技术来对以上问题进行优化，以此搭建一个满足回应型政府要求的政务问答系统。

近年来，LLMs 已经重塑了人工智能研究的格局。LLMs 有强大的文本推理能力，在政务问答领域具有极佳的应用优势。Hua[25]设计了一种基于 ERNIE+CNN 的智能问答系统，通过知识模型和语义分析技术，提高了政策问题回答的准确性。Cui[26]结合 ALBERT 和胶囊网络，提出了一种有效的政府公文关系抽取方法，提升了模型性能和关系分类效果并解决了政府公文数据集不足的问题。以上研究成功将 LLMs 应用于政务问答领域，取得显著的性能提升，但仍存在一定的局限。首先，以上研究训练成本较高，对训练数据的质量和数量都有较高要求，很难将其迁移应用至其他场景的政务数据源。其次，以上研究并没有使用最新的 LLMs 且采用的评估指标由于数据集结构限制不能较好地评估模型文本生成的能力，其文本生成性能存在一定的提升空间。在系统评估环节仅关注模型指标性能的提升，而 LLMs 是文本生成模型，其生成内容的有效性同样也是也需要进行详细评估，而且这很大程度地影响到系统是否符合当下回应型政府的要求。因此，本研究将采用最新的 LLMs 模型作为政务问答系统的基底模型，除了关注模型指标提升外，系统生成文本的有效性和可用性也会被纳入讨论当中，以此达到搭建回应型政府服务的研究目标。

2 数据概况

本研究使用的数据集为中国人民网政务留言板的问答数据。该数据集包含了市民与人民网平台之间的问答文本数据，即市民通过留言板提交政务咨询问题，工作人员在政务平台上对这些问题进行针对性回复。该数据集的总量为 130,258，涵盖了 2022 年全年全国各省的政务留言板问答数据。除此之外，它同时包含了问题时间、问题类型（城建、环保、交通、教育、金融、就业、旅游、企业、三农、文娱、医疗、治安）、问题所在的省市、用户点赞数等字段信息。这些字段信息反映了市民问题的多样化信息，有助于后续系统搭建对问答数据进行针对性筛选和检索。选择这一年份的数据主要是基于数据量与各类型数据平衡的考量。该年份是近五年政务问题量最大的一年，能为模型训练和测试提供足够的样本数据。此外，该年份的政务问题数据在不同类型及其在各个时间段的分布上呈现出较高的均衡性，这有助于避免模型对特定类型的问题或特定时间段的数据产生过度拟合的现象。因此，本研究的系统搭建是基于该数据集进行训练和测试的。

3 系统框架

本研究构建的 GovLLM 系统如图 1 所示，主要有两部分组成，第一部分为 LLMs 微调，系统以阿里云开发的 Qwen-14B 为基底 LLMs，对中文文本具有更出色的理解能力。LLMs 基于训练集进行微调，使其具备理解政务咨询语境的能力。当有新的政务咨询问题输入时，经过微调的 LLMs 接受新问题并将新问题传递到第二部分。第二部分为基于多智能体（Multi-Agent）的增强检索。训练集除了用于第一部分的 LLMs 微调之外，在第二部分中用于构建基于 BERT 生成的历史问答向量检索库。BERT（Bidirectional Encoder Representations from Transformers）在构建历史问答向量检索库中扮演着核心角色，它通过其深度双向 Transformer 架构，能够理解并生成高质量的上下文相关向量，这些向量能够精确地表示和匹配历史问答数据，从而实现高效准确的检索功能。而第一部分中输入的政务咨询问题通过多个智能体（总结 Agent、分类 Agent）的提取和筛选得到历史同地区最新且回答收获点赞数较高的相似问题，最终通过政务工作人员 Agent 结合新问题和历史相似问题回答生成对于新问题的参考回答，以此服务于政务咨询的快速有效回复。

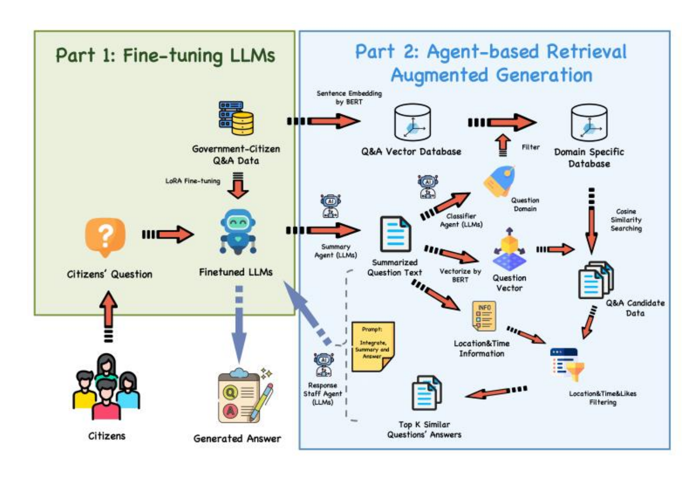


图 1 GovLLM 框架图

Fig. 1 Framework of GovLLM

3.1 LLMs 微调

尽管本研究采用的基底 LLMs 在理解通用中文文本的任务上表现出色，且积累了广泛的语料知识，但在特定领域和下游任务中仍有显著提升空间。尤其是在政务咨询领域，LLMs 需要对政务文本具有更加深入的情境理解能力和特定领域表达。为此，本研究采用微调技术来调整基底 LLMs 的参数，以适应与政务咨询问答数据集。微调能够使基底 LLMs 更好地理解政务咨询领域的知识，从而有效满足实际应用需求。

由于 LLMs 参数量庞大，全参数微调会多批次消耗大量计算资源。因此，本研究采用一种高效的参数微调方法——低秩自适应[27]（Low-Rank Adaptation, LoRA）——以适应政务咨询领域的任务，同时减少资源消耗。LoRA 技术的核心在于冻结预训练 LLMs 的参数权重，并引入两个矩阵 A 和 B 来替代微调中改变的参数。在微调过程中，仅更新 A 和 B 矩阵，而不是全

部模型参数。这种方法通过低秩分解模拟参数变化，以极小的参数增量实现对 LLMs 的间接训练，从而在现有资源下更好地适应政务咨询回复的任务。经过 LoRA 微调后，基底 LLMs 用于后续构建智能体和文本生成的任务。

3.2 多智能体的检索增强

GovLLM 系统的第二部分主要是基于多智能体对历史相似问题进行增强检索。相较于直接的检索增强，基于多智能体的增强检索不仅可以提取主要文本信息，减少噪音，而且能够通过协同运行共享关键信息，最大程度地保留政务咨询问题的上下文信息。这能够让历史问题检索更加准确，使得 LLMs 的推理和生成内容更加有效。

3.2.1 历史问答向量数据库搭建

为了使历史政务问答文本数据更易于检索和匹配，本研究将训练集中的历史政务问题文本转化成文本向量索引库，即应用 BERT 预训练模型对文本进行语义编码，以句向量的形式存储每一个历史政务咨询问题文本。该向量索引库可用于筛选和查询，它既可以根据问题类别、回答点赞数进行筛选得到特定的子向量库，也能通过计算新问题向量与向量数据库中存储的向量之间的相似度,可以快速检索到历史相似问题。

3.2.2 Agent 设计

GovLLM 的智能体系统由三个核心 Agent 组成，每个 Agent 承担不同的任务，以此提高文本推理和检索的效率和准确性。这三个 Agent 的功能都是通过基底 LLMs 每次对话中的 Role 和 Content 参数进行设定。

总结 Agent： 总结 Agent 的主要功能是提取和总结市民政务咨询问题的关键信息。通过分析咨询文本，该 Agent 能够识别并提取问题背景、问题现状以及核心诉求三个核心要素。它类似于政务平台中负责数据清洗的工作人员，不仅能够简化问题，减少不同市民表达方式带来的信息偏差，还能为后续的处理步骤提供精确和简化的问题描述。总结 Agent 的设计目的是通过精准的信息提取，为分类和响应生成奠定基础。

分类 Agent： 分类 Agent 负责将总结过后的政务咨询问题归类到预设的咨询类别中。该 Agent 通过 Prompt 的形式将问题自动分配到城建、环保、交通、教育、金融、就业、旅游、企业、三农、文娱、医疗、治安、民生其中之一的类别当中。分类 Agent 类似于政务平台中负责问题归类和分发的工作人员，其设计目的是通过高效的分类机制，将每个输入的问题进行正确分类，并且基于问题类别筛选得到特定类别问题的历史数据库，从而增强检索的针对性和搜索效率。

政务答复 Agent： 政务答复 Agent 是基于总结和分类检索的结果，生成针对性回复的关键组件。该 Agent 会假定自身为负责政务答复的工作人员，综合考虑总结 Agent 提供的问题核心要素以及分类 Agent 的分类结果，同时检索历史相似同地区最新问题及其高赞回复，以此为基础生成针对性的回复。政务答复 Agent 的设计目的是模拟政务工作人员答复市民咨询的过程，借助基底 LLMs 的文本理解和推理能力，减少人工介入，提高响应速度和市民满意度。

表 1 Agent 提示词模板

Table 1 Agent prompt templates

Agent 类型	提示词模板
总结 Agent	你是一个出色的政务问题总结专家，你在政务平台中负责提取和总结市民政务问题中的要点并进行数据清洗，你将收到一个政务问题：\n 问题：**{政务问题原始文本}**\n 请你从该问题中提取问题背景、问题现状以及核心诉求三个核心要素，并以 json 文件格式的工单列表返回。
分类 Agent	你是一个出色的政务问题分类专家，你在政务平台中负责将总结后的政务问题归类到预设的问题类别中，你将收到一个总结后的政务问题：\n 问题：**{政务问题 json 文本}**\n 请你将该问题将问题归类到到城建、环保、交通、教育、金融、就业、旅游、企业、三农、文娱、医疗、治安、民生其中之一的类别，并将该类别直接返回。
政务答复 Agent	你是一个出色的政务问题答复专家，你在政务平台中负责政务问题的答复工作，你将收到一个发生在**{问题位置省市}**的总结后的政务问题以及与该问题密高度相似的同地区较新的历史问题的高赞答复：\n 问题：**{政务问题 json 文本}**\n Top K 历史相似问题回复：**{[历史回复文本列表]}**\n 请你基于以上历史相似问题的回复，对新的政务问题进行答复并进行返回。

三个 Agent 对应的提示词模板如表 1 所示。这三个 Agent 彼此协同运作，搭建起一个完整的信息传递链条，共同构成了一个高效、智能的政务咨询处理系统，旨在通过自动化和智能化手段，提升政务咨询的响应质量和效率。

3.2.3 基于向量相似度和问题属性的增强检索

基于已经构建好的历史问题向量索引库和三个核心 Agent，GovLLM 系统从历史数据中高效且准确地检索同地区、最新、最相似且回答点赞数较高的历史政务咨询问题。具体如下：

首先，总结 Agent 对新输入的政务咨询问题进行分析，提取关键信息，简化文本并生成问题向量。同时，分类 Agent 也会对输入问题进行归类，得到问题分类后根据问题类别进一步筛选历史问题索引库的同类别子库，以缩小检索范围和提高检索效率。在此基础上，系统采用余弦相似度匹配算法对子库中的历史问题向量进行粗筛，以识别与新问题相似度较高的候选问题供下一步细筛。

接下来，系统根据新问题的位置和时间信息，对粗筛结果进行精筛和排序，以确保所选问题不仅在地理和时间上与新问题相近，而且市民对于回复的满意度也较高。这一步骤的筛选标准为：同省市、同类别问题、时间新、点赞数高，从而得到排序靠前的精排结果，以此得到 Top K（K 可进行自定义，默认为 5）个经过精筛得的历史相似问题及其回复。

最后，政务答复 Agent 结合总结 Agent 的提炼结果和精排筛选的历史相似问题回复，生成参考回复，以供政务工作人员进行回复时参考和修改。这一过程不仅提高了问题处理的自动化水平，还确保了回复的有效性、针对性和时效性，从而提升了政务咨询的响应质量和效率。

通过这一系统设计，GovLLM 系统能够有效地利用历史数据，为政务咨询问题提供智能化的处理方案，以此达到建设回应型政府的研究目标。

3.3 系统评估

BERTScore 通过利用 BERT 模型的预训练上下文嵌入，并通过余弦相似度来衡量生成文本与参考文本中单词的匹配程度[28]。该过程首先将生成文本和参考文本转换成嵌入表示，随后利用 BERT 模型评估这些嵌入之间的相似度分数，从而反映出两段文本间的语义相似度。其中，R_{BERT}（Recall BERT）衡量的是参考文本与生成文本之间的最大相似度之和除以参考文本总数，代表召回率；P_{BERT}（Precision BERT）衡量的是生成文本与参考文本之间的最大相似度之和除以生成文本总数，代表精确度；F_{BERT}（F1 BERT）则是精确度和召回率的调和平均值，它结合了精确度和召回率，以提供一个平衡的相似度评估指标，相关指标如公式(1)～(3)所示：

$$R_{BERT}=\frac{1}{|x|}\sum_{x_i\in x} max_{y_i\in y}^{x_i^T y_j} \tag{1}$$

$$P_{BERT}=\frac{1}{|y|}\sum_{y_j\in y} max_{x_i\in x}^{x_i^T y_j} \tag{2}$$

$$F_{BERT}=2\times\frac{P_{BERT}\times R_{BERT}}{P_{BERT}+R_{BERT}} \tag{3}$$

其中，x 表示用于评估生成文本质量的参考文本，x_i 表示第 i 个参考文本，y 表示系统生成文本，y_j 表示第 j 个系统生成文本。

相比于传统的评估指标如 BLEU 和 METEOR，BERTScore 具有显著优势。它能够捕捉到底层的词间关系和语义信息，而不只是简单的词汇排列组合，这使得 BERTScore 在评估文本的语义相似性方面更为有效。BERTScore 与人类判断具有更好的相关性，其评估结果更接近人类对文本相似度的直观感受，它不那么容易受到常用词干扰，因为它不仅仅依赖于 n-gram 的匹配机制，而是考虑了整个句子的上下文信息。综合这些优势，BERTScore 作为本研究评估系统生成文本质量的主要指标。

4 实验结果及分析

4.1 测试数据

本研究选取总数据集中回答点赞数排名前 1000 的问答对作为测试集，并将测试集从训练集中剔除。测试集中的问题作为输入以获得系统的生成文本，测试集中的回答作为评估系统生成文本质量的参考文本。本研究通过计算系统生成文本与参考文本的 BERTScore 来衡量和量化生成文本的质量，从而评估系统回答政务问题的能力高低。

4.2 基准参考模型

为了评估 GovLLM 的性能，本研究选取了几个常见的基线模型进行了比较，包括 GPT-4o、Qwen-14B、Kimi 和 ERNIE。GPT-4o 是由 OpenAI 开发的最新大语言模型。Kimi 是由 Moonshot AI 开发的最新多语言对话系统，而 ERNIE 则是百度推出的一款最新的的语义理解模型。这些模型共同构成了本研究的基准模型，有助于准确衡量 GovLLM 的性能。

4.3 生成文本质量测试

本研究基于上述测试数据集对 GovLLM 和其他基线模型进行生成文本质量测试，主要衡量指标是这些模型基于测试集的平均 BERTScore。测试结果包含模型在全量数据集的整体表现以及对于部分类别政务问题的表现，结果如表 2 所示。

表 2 生成文本质量实验结果

Table 2 Experimental results on generated text quality

测试数据	指标	GPT-4o	Qwen-14B	Kimi	ERNIE	GovLLM
全量	R_{BERT}	0.8239	0.7294	0.7534	0.7605	**0.9349**
	P_{BERT}	0.8207	0.7256	0.7630	0.7612	**0.9286**
	F_{BERT}	0.8223	0.7275	0.7582	0.7608	**0.9317**
三农类	R_{BERT}	0.8436	0.7362	0.7521	0.7931	**0.9073**
	P_{BERT}	0.8491	0.7474	0.7518	0.8064	**0.8982**
	F_{BERT}	0.8463	0.7418	0.7519	0.7997	**0.9027**
教育类	R_{BERT}	0.8187	0.7362	0.7479	0.7534	**0.9156**
	P_{BERT}	0.8295	0.7474	0.7408	0.7590	**0.9065**
	F_{BERT}	0.8241	0.7418	0.7443	0.7562	**0.9110**
交通类	R_{BERT}	0.8273	0.7254	0.7164	0.7472	**0.9390**
	P_{BERT}	0.8168	0.7210	0.7105	0.7461	**0.9321**
	F_{BERT}	0.8220	0.7232	0.7134	0.7466	**0.9355**

从表 2 所示结果可以看出，GovLLM 在全量数据集中的表现优于四个基线模型，且有 10%左右的性能提升，这意味着 GovLLM 生成的文本与高赞回复文本高度相似，具有较高的文本质量和应用价值。另外，在特定问题类型上，GovLLM 也保

持了在全量数据集上的同等表现，而部分基线模型在特定类型的测试集在存在较大的性能波动。得益于 GovLLM 中分类 Agent 的设计，GovLLM 系统能够从历史数据中针对性地检索与测试问题相同类型的历史问题及其答复，并以此生成面向该类型问题的针对性回复，这样的设计使得 GovLLM 在不同类型的政务问题都能保持稳定且优异的生成文本质量。

4.4　参数敏感性测试

虽然 GovLLM 系统的整体框架是确定的，但其中仍包含大量可调整的参数设置。而这些参数的不同设置方案会很大程度地影响 GovLLM 的效率与性能。同时，实际应用场景与实验场景存在较大差别，实际应用场景往往不具备足够的算力资源。因此，为了测试 GovLLM 对于这些参数的敏感性并寻找最合适实际应用场景的参数调整方案，本研究对 GovLLM 中的 Top K 数（输入政务答复 Agent 里历史相似回复的数量）、微调数据量、历史向量库数据量进行控制变量测试，并以 F_{BERT} 作为评估变量，结果如图。

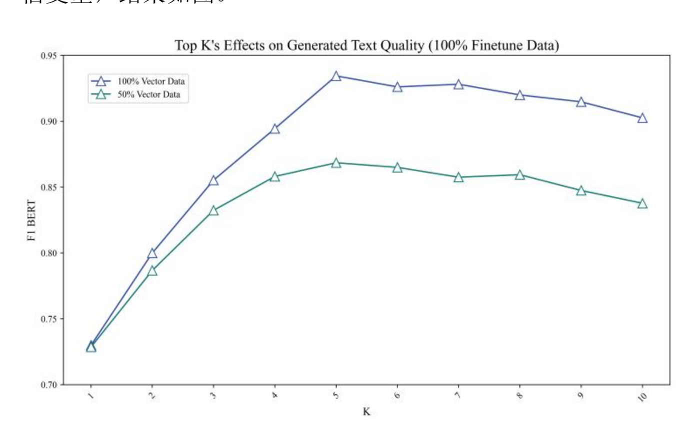


图 2　Top K 数参数敏感性测试

Fig. 2　Top K sensitivity test

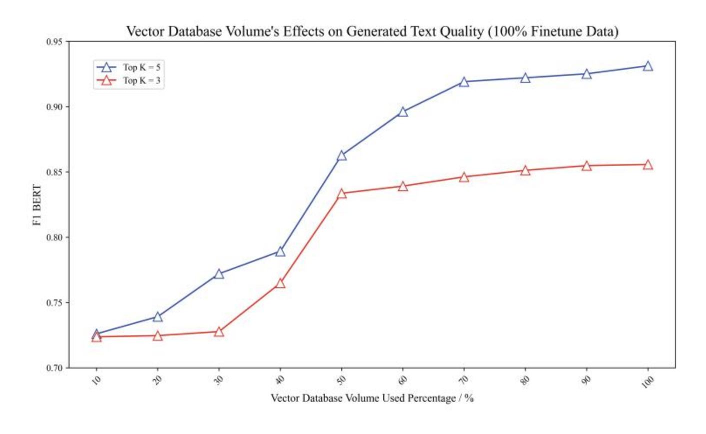


图 3　向量数据库大小参数敏感性测试

Fig. 3　Vector database volume sensitivity test



图 4　微调训练数据大小参数敏感性测试

Fig. 4　Finetune training data sensitivity test

图 2 展示了在全量微调的情况下，Top K 值的增加在 K 值较小时显著提升了 F_{BERT} 值，从而提高生成文本质量，但当 K 值超过 5 后，F_{BERT} 值呈现轻微下降趋势。这种效应对于全量使用历史向量数据的情况会更明显。可能的解释是，当提供给政务答复 Agent 的相似回复较少时，这些回复与新问题的相关性高且质量较好，Agent 能快速学习相似回答的规律并生成高质量答复。然而，当 K 值超过一定范围，相似回复的相关性和质量下降，导致 Agent 从低质量回复中学习到噪声信息，从而降低了答复质量。因此，在实际应用中，应通过敏感性测试确定 K 值的最优阈值，例如 GovLLM 系统中 K 值可设为 5。

图 3 探讨了全量微调下历史向量库数据使用量对生成文本质量的影响。实验发现历史向量库数据使用量与 F_{BERT} 值的关系非线性。当向量数据库数据使用量较少时，F_{BERT} 值随使用量增加而显著提升，但超过一定阈值后，增长速度减缓，最终趋于稳定。另外，K 值的选择也影响 F_{BERT} 值，K=5 相较于 K=3 能更好地利用历史向量数据库，提高文本生成质量。实际应用中，尽管全量使用历史向量数据库能充分利用数据，但也会增加检索时间，降低效率。对于 GovLLM 系统，在全量微调和 Top K=5 的条件下，选择 70%的向量数据用量可实现 90%以上的 F_{BERT} 值，同时减少检索时间，可满足算力资源有限但对文本质量有一定要求的应用场景。

图 4 分析了微调数据量对生成文本质量的影响。微调数据量与 F_{BERT} 值呈正相关，但提升效果有限，增长速度较慢。相比之下，较高历史向量数据库的用量能显著提升 F_{BERT} 值。在实际应用中，LLMs 微调有助于模型快速理解政务领域语境和表达模式，提高文本质量。然而，大量微调数据训练的边际效益较低，需根据算力资源和效率需求合理分配微调数据和向量库数据比例，以最大化系统性能。

4.5　消融实验

为了检验 GovLLM 中微调模块、向量数据库模块、Agent 模块、基于问题属性检索排序模块对于整体系统的性能提升贡献及有效性，本研究将 GovLLM 中的模块进行部分去除形成五个新的模型，并基于全量测试集进行生成文本质量测试，结果如表 3 所示。表 3 中，-RFT (remove finetune) 指的是去除微调

模块，-RVD(remove vector database)指的是去除向量数据库，-RA (remove agent) 指的是去除所有的 Agent 模块(用直接的问答提示词替代)，-RR(remove ranking)指的是去除基于问题属性和回复属性的排序模块(仅以相似度匹配召回历史相似回复)。

表 3　消融实验结果

Table 3　Ablation experiment results

消融模型	添加模块	R_{BERT}	P_{BERT}	F_{BERT}	$\triangle F_{BERT}$
GovLLM-RR-RA-RVD-RFT	-	0.7294	0.7256	0.7275	-
GovLLM-RR-RA-RVD	FT	0.7546	0.7524	0.7535	2.6154%
GovLLM-RR-RA	VD	0.8073	0.8038	0.8055	5.2048%
GovLLM-RR	A	0.8630	0.8712	0.8671	**6.1534%**
GovLLM	R	0.9349	0.9286	0.93173	**6.4659%**

表 3 展示了消融实验结果，该实验测试并记录了基底模型逐步添加新模块后的指标及其变化值。从结果来看，微调训练虽然能够让 LLMs 理解政务领域的语料，但是对生成文本质量的提升效果较为有限。历史向量数据库模块的引入较显著地提高了模型的 F1 BERT 值，因为历史向量数据库显式地通过提示词方式引导并监督 LLMs 的文本生成，其提升效果显著于微调训练。Agent 模块和基于问题回复属性检索排序模块对于 LLMs 的提升是所有模块中最显著的。Agent 模块通过提取关键信息、确定问题类型、缩小检索范围，并整合历史相似的高质量回复，从而提高系统生成文本的质量。属性检索排序模块通过筛选出保留同类别同地区的相似问题高质量回复，排除不相关数据，确保回复内容的地区适用性和有效性。综上所述，GovLLM 的四个模块各有功能和效用，既能为基底模型带来不同程度的性能提升，提高生成文本质量，也能提高检索生成效率，增强文本内容在政务领域的有效性和实用性。

4.6　结果分析与讨论

表 4 展示了 GovLLM 与其他基线模型对于同一个输入问题的生成回复文本。从结果可以看出来，基线模型并没有捕捉到广州这一重要的位置信息，因此基线模型的回复虽然表面上回答了问题，但其内容实际上并不能直接服务于广州市民，使其回复内容的实用性和有效性降低，这也是目前政务问答系统需要提升和优化的地方。GovLLM 得益于历史向量数据库和智能体系统等模块的协同作用，生成的回复文本不仅与广州市的实际情况相符，还能提供切实可行的解决方案，显著提升了回复的有效性和实用性。市民可按其内容解决该政务问题，政务平台也可基于生成的参考回复进行编辑修改，以形成最终的回复，提高了政务平台回复问题的效率。

表 4　系统生成回复文本示例

Table 4　System-generated answer text examples

模型	输入问题：社保账户余额数目对不上，应该如何处理？（广东省广州市）
GPT-4o	你可以自查个人社保缴纳记录，确认是否有遗漏或错误。如果自查无误，可以采取以下步骤：1. 登录社保官方网站或使用官方 APP，查看是否有系统公告说明正在进行数据维护或更新，这可能是导致数据不一致的原因。2. 联系你的工作单位，确认他们是否按时足额为你缴纳社保。
Qwen	如果你的社保账户余额数目对不上，可能是系统出错了。你可以尝试重启电脑或者重新登录系统。如果还是显示异常，只能暂时先等系统自身进行修复。
Kimi	社保账户余额数目对不上可能是由于系统更新延迟、数据录入错误或者账户信息有误等原因造成的。你可以首先检查自己的社保缴纳记录，确认是否有遗漏或者错误。
ERINE	社保账户余额数目对不上，可能是系统更新延迟或者数据录入错误。你可以尝试联系社保局，他们应该能帮你解决这个问题。记得带上你的身份证和社保卡，这样他们能更快地帮你处理。
GovLLM	面对社保账户余额数目对不上的问题，以下是适用于广东省广州市地区的解决措施： ● 自查缴费记录：首先，您可以通过**“粤省事”微信小程序**进行查询。登录小程序后，选择“社保”服务，即可查看您的社保缴费记录，确认是否有遗漏或错误。 ● 核实单位缴纳情况：联系您的工作单位，确认他们是否按时足额为您缴纳社保。 ● 检查系统公告：登录**广州市劳动保障网上办事系统**或使用“粤省事”小程序，查看是否有系统公告说明正在进行数据维护或更新，这可能是导致数据不一致的原因。 ● 联系社保服务中心：如果自查和核实后问题仍未解决，您可以携带个人身份证明文件，前往**广州市劳动与社会保障局**或通过**社保热线 12333** 进行咨询，请求工作人员协助核查账户信息，并解决余额数目不符的问题。 ● 提交书面申请：如果需要，您可以提交书面申请，要求社保局进行账户核查和调整。 ● 跟进处理进度：在提交申请后，定期跟进处理进度，确保问题得到及时解决。 ● 法律途径：如果问题复杂且长时间未能解决，您可以考虑咨询法律专业人士，了解可能的法律途径。 请记得，社保账户的准确性关系到个人权益，因此及时处理和解决余额数目不符的问题是非常重要的。通过上述步骤，您可以更有效地解决社保账户余额数目对不上的问题。

基于已有实验及结果，本研究得出了一系列有意义的发现，并从算法和组织管理两个层面进行讨论。

从算法层面，本研究为 LLMs 在政务领域的应用提供了实际其可扩展的方案。首先，研究的模块设计（向量数据库、智能体等）充分发挥了 LLMs 的模式识别和文本生成能力，其生成的内容变得更加可控和有效，以便于相关人员进行针对性地调整和修改，实现跨领域的应用。其次，参数敏感测试量化了系统中不同参数选择对最终系统性能的影响，有利于相关人员根据测试结果和实际的算力资源选择合适的参数，最终应用于实际场景当中。最后，消融实验量化了不同模块的提升效果，有利于相关人员将有限资源优先优化系统中的特定模块，最大限度地提高系统性能。综上所述，本研究不仅提供了一个可应用于政务问题回复的系统框架，而且其相关实验结果也能有助于其实际落地应用。

从组织管理层面，本研究为建设"回应型政府"的目标提供了一个有效的工具系统。首先，研究提出的 GovLLM 系统有助于提高政务平台的大数据处理能力，尤其是对于数量庞大的政务问题的回复效率。工作人员不需要把时间精力花费在繁琐的查找检索工作，而只需要根据生成的参考回复进行校验和修改即可，提高了回复效率并减少了不必要的资源投入。其次，基于智能体系统和检索排序模块，GovLLM 可以在单系统内集成全国各省市的政务问题回答任务，这有利于消除各省市的政务服务信息差，搭建"一网通办"的政务服务平台，为建设全国范围的回应型政府提供示例方案。对于市民而言，市民无需繁琐地查询不同省市的政务问答平台，只需要在单平台便能实现全国不同省市的政务问答。对于政府而言，管理和维护统一化的政务平台会更高效，且各省市之间的政策信息差也会很大程度地减少。最后，本研究也促进了各政务平台建设"回应型政府"的进程。本系统不仅能够帮助政务平台及时高效地回复大量的政务问题，也提高了回复的有效性和实用性，充分考虑了我国不同地区的实际情况差异，且回复内容对于市民而言更加具体和符合当地实际情况。这与我国建设"回应型政府"的核心理念是相一致的。

5 结束语

本研究成功构建了 GovLLM 系统，其为一个基于 LLMs 的政务参考回复生成系统，相较于基线模型提升了模型表现 5-11%，有效扩宽了 LLMs 在政务领域的应用能力。GovLLM 系统通过集成了历史向量数据库、智能体系统、检索排序等多个模块，提高了回复的效率和文本质量，确保了内容与地区实际相符，缓解了 LLMs 文本生成的幻觉，增强了系统的可用性。此外，该系统有助于降低了纯人工回复的工作成本，提升了政务平台的运作效率。总体而言，GovLLM 系统为建设"回应型政府"提供了实际有效的技术方案，推动了回应型政务平台的智能化进程。

6 参考文献

[1] Alvarenga A, Matos F, Godina R, CO Matias J. Digital transformation and knowledge management in the public sector[J]. Sustainability, 2020, 12(14):5824.

[2] Gamper J, Augsten N. The role of web services in digital government. [C]//Proc of Conference on Electronic Government. Berlin: Springer, 2003: 161-166.

[3] Bélanger F, Carter L. The impact of the digital divide on e-government use[J]. Communications of the ACM, 2009, 52(4):132-135.

[4] Lee-Geiller S, Lee TD. Using government websites to enhance democratic e-governance: a conceptual model for evaluation[J]. Government Information Quarterly, 2019, 36(2):208-225.

[5] Bastos D, Fernández-Caballero A, Pereira A, Rocha NP. Smart city applications to promote citizen participation in city management and governance: a systematic review[J]. [5], 2022, 9:89-118.

[6] 陈文权,余雅洁.网络环境下服务型政府建设的回应性及路径研究——以2013年五省(市)书记和省长集中回复网友留言为例[J].中国行政管理, 2014, (7):74-77. (Chen Wenquan, Yu Yujie. Research on the responsiveness and path of service-oriented government construction in the network environment: taking the 2013 consecutive replies of secretaries and governors of five provinces (cities) to netizens' messages as an example [J]. Chinese Public Administration, 2014, (7): 74-77.)

[7] Gupta P, Gupta V. A survey of text question answering techniques[J]. International Journal of Computer Applications, 2012, 53(4):1-8.

[8] Daniel J, James H. Speech and language processing[M]. 2nd ed. Upper Saddle River: Prentice Hall, 2014.

[9] Hao Tianyong, Li Xinxin, He Yulan, Wang Fu Lee, Qu Yingying. Recent progress in leveraging deep learning methods for question answering[J]. Neural Computing and Applications, 2022, 1:1-9.

[10] Dong Hui, Yu Siwei, Jiang Ying. Text mining on semi-structured e-government digital archives of China[C]//Proc of 2009 Second Pacific-Asia Conference on Web Mining and Web-based Application. Piscataway, NJ: IEEE Press, 2009: 11-14.

[11] Gao Shangsheng, Gao Li, Li Qi, Xu Jianjun. Application of large language model in intelligent q&a of digital government[C]//Proc of the 2023 2nd International Conference on Networks, Communications and Information Technology. New York: ACM Press, 2023: 24-27.

[12] Nigam SK, Mishra SK, Mishra AK, Shallum N, Bhattacharya A. Legal question-Answering in the Indian context: efficacy, challenges, and potential of modern AI model[OL]. (2023-09-26)[2024-11-06]. https://arxiv.org/abs/2309.14735.

[13] Tang Ruixiang, Yu-Neng Chuang, Xia Hu. The science of detecting llm-generated text[J]. Communications of the ACM, 2024, 67(4):50-59.

[14] Wu Shijie, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, Kambadur P, Rosenberg D, Mann G. Bloomberggpt: a large language model for finance[OL]. (2023-03-30)[2024-11-06]. https://arxiv.org/abs/2303.17564.

[15] Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS.

Large language models in medicine[J]. Nature Medicine, 2023, 29(8):1930-1940.

[16] Baek J, Jeong S, Kang M, Park JC, Hwang SJ. Knowledge-augmented language model verification[OL]. (2023-10-19)[2024-11-06]. https://arxiv.org/abs/2310.12836.

[17] Tan Bowen, Yang Zichao, AI-Shedivat M, Xing EP, Hu Zhiting. Progressive generation of long text with pretrained language models[OL]. (2020-01-28)[2024-11-06]. https://arxiv.org/abs/2006.15720.

[18] Welch EW, Hinnant CC, Moon MJ. Linking citizen satisfaction with e-government and trust in government[J]. Journal of Public Administration Research and Theory, 2005, 15(3):371-391.

[19] Katsonis M, Botros A. Digital government: a primer and professional perspectives[J]. Australian Journal of Public Administration, 2015, 74(1):42-52.

[20] Zhang Yuhao. Digital Government Construction Perspective: a Study on the path to promote the improvement of government public management capacity[J]. Frontiers in Business, Economics and Management, 2022, 6(3):35-40.

[21] Patsoulis G, Promikyridis R, Tambouris E. Integration of chatbots with Knowledge Graphs in eGovernment: the case of getting a passport[C]//Proc of the 25th Pan-Hellenic Conference on Informatics. New York: ACM Press, 2021:425-429.

[22] Schwarzer M, Düver J, Ploch D, Lommatzsch A. An Interactive e-Government Question Answering System[J]. Labor & Workforce Development Agency, 2016:74-82.

[23] Beltrán A, Ordoñez S, Monroy S, Melo L, Duarte N. Question-answering systems in the specific domain of e-government[OL]. (2016-04-01)[2024-11-06]. http://dx.doi.org/10.13140/RG.2.1.2491.6889.

[24] Rodsawang C, Thongkliang P, Intawong T, Sonong A, Thitiwatthana Y, Chottanapund S. Designing a competent chatbot to counter the COVID-19 pandemic and empower risk communication in an emergency response system[J]. Outbreak, Surveillance, Investigation & Response (OSIR) Journal, 2020, 13(2):71-77.

[25] 华斌, 康月, 范林昊. 政策文本的知识建模与关联问答研究[J]. 数据分析知识发现, 2022, 6(11):79-92. (Hua Bin, Kang Yue, Fan Linhan. Knowledge modeling and association Q&A for policy texts[J]. Data Analysis and Knowledge Discovery, 2022, 6(11):79-92.)

[26] 崔从敏, 施运梅, 袁博. 面向政府公文的关系抽取方法研究[J]. 计算机技与发展, 2021, 31(12):26-32. (Cui Congmin, Shi Yunmei, Yuan Bo. Research on relation extraction method for government documents[J]. Computer Technology and Development, 2021, 31(12):26-32.)

[27] Hu EJ, Shen Yelong, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang Lu, Chen Weizhu. Lora: Low-rank adaptation of large language models[OL]. (2021-01-17)[2024-11-06]. https://arxiv.org/abs/2106.09685.

[28] Zhang Tianyi, Kishore V, Wu Felix, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert[OL]. (2019-04-21)[2024-11-06]. https://arxiv.org/abs/1904.09675.